DOCUMENT RESUME

ED 357 353 CS 213 806

AUTHOR Lai, Morris K.; Saka, Thomas
TITLE Using Differential Item Functioning Procedures To Improve Interpretation of and Performance on the Verbal Subtest of the SAT.
PUB DATE 14 Apr 93
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; High Schools; High School Seniors; Instructional Effectiveness; Public Schools; *Scores; *Test Items; Test Theory; Test Wiseness; *Verbal Tests
IDENTIFIERS Hawaii; Scholastic Aptitude Test

ABSTRACT

Two studies investigated factors affecting the scores of Hawaii students taking the verbal subtest of the Scholastic Aptitude Test (SAT). For the past several years, the mean verbal scores of Hawaii students have consistently been among the lowest 10% of all states. The first study addressed the identification of items and types of items that have been answered differentially by Hawaii students in comparison to mainland United States students. The items were identified through differential item functioning (DIF) procedures which assess performance differences between groups of individuals with the same overall scholastic aptitude. Results indicated that Hawaii students performed less well than the mainland reference group (students of equal overall scholastic aptitude) on the early items in each of the antonym sections and better than the reference group on the more difficult or later items in each section. Carelessness and unfamiliarity with the item type were identified as possible causes. The second study utilized a sample of Hawaii public school students who received brief instruction addressing the low performance on the types of items identified in the first study. Independent groups t-tests conducted between the treatment students and a sample from the original pool of 1988 examinees showed equivalence of performance on the pretest. Treatment students performed statistically significantly better than the comparison groups did after receiving the treatment. (Five tables of data are included. Contains 17 references.) (Author/RS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED357353

Using Differential Item Functioning Procedures to Improve Interpretation of and Performance on the Verbal Subtest of the SAT

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Morris K. Lai

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Morris K. Lai
University of Hawai'i

Thomas Saka
Hawai'i State Department of Education

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

CS213806

Paper presented at the annual meeting of the American Educational Research Association, Atlanta
Session 25.17, April 14, 1993

ABSTRACT

This study investigated factors affecting the scores of Hawai'i students taking the verbal subtest of the *Scholastic Aptitude Test* (*SAT*). For the past several years, the mean verbal scores of Hawai'i students have consistently been among the lowest 10% of all states.

The investigation consisted of two major studies. Study 1 addressed the identification of items and types of items that have been answered differentially by Hawai'i students in comparison to Mainland U.S. students. We identified the items through differential item functioning (DIF) procedures which assess performance differences between groups of individuals with the same overall scholastic aptitude. The results of the analyses indicated that Hawai'i students performed less well than the Mainland reference group (students of equal overall scholastic aptitude) on the early items in each of the antonym sections and better than the reference group on the more difficult or later items in each section. Carelessness and unfamiliarity with the item type were identified as possible causes.

Study 2 utilized a sample of Hawai'i public school students who received brief instruction addressing the low performance on the types of items identified in Study 1. Independent groups *t*-tests conducted between the treatment students and a sample from the original pool of 1988 examinees showed equivalence of performance on the pretest. Treatment students performed statistically significantly better at the .05 level than the comparison groups did after receiving the treatment.

Introduction

Recent research studies have identified a number of factors related to differential item functioning (students of equivalent scholastic aptitude performing differently on specific test items) on the College Board's *Scholastic Aptitude Test* across different groups of examinees. Most studies have focused on Black-White differences (Hackett, Holland, Pearlman, & Thayer, 1987; Freedle & Kostin, 1987; Lord, 1977). Other studies have concentrated on differences between Whites and Asians (Wright, 1983), Whites and Hispanics (Schmitt, 1985), and males and females (Dorans & Kulick, 1983; Lawrence, Curley, & McHale, 1988).

The previously cited studies have concluded that the differential performance can be attributed to factors linked to the content of the items. The factors identified include the presence of technical (as opposed to philosophical) and science material, abstractness of the individual terms or relations, declarative knowledge base (vocabulary and general knowledge), language skills, and semantic relationships (e.g., class inclusion, part-whole, cause/purpose). Additional factors involve specific item characteristics such as length of the passage or type of information being requested in reading comprehension items and the number of blanks in sentence-completion items.

For the past several years, the mean scores of Hawai'i students on the verbal subtest of the *SAT* have been among the lowest 10% of all states. To investigate how differential item functioning (DIF) may be related to the performance of Hawai'i students on the verbal subtest of the *SAT*, a two-part research design was used. The first part addressed the identification of items and types of items that have been answered differentially by Hawai'i students in comparison to Mainland U.S. students of equivalent scholastic aptitude. In the second part of the investigation a related instructional treatment was developed, administered, and tested for its effectiveness.

Study 1: Determination of DIF Between Hawai'i and Mainland Students

Sample

The participants used in Study 1, dealing with the determination of DIF between Hawai'i and Mainland students, were a randomly selected and representative sample of 114,029 college bound seniors who were administered the *Scholastic Aptitude Test* nationwide in November 1988.

Instrumentation and Analysis of the Data

We applied the Mantel-Haenszel procedure (Holland & Thayer, 1988) to the 85 verbal subtest items consisting of antonym, sentence-completion, reading-comprehension, and analogy items to detect test items which were functioning differentially between the Hawai'i and Mainland examinees. This matched group comparison procedure, originally developed by Mantel and Haenszel (1959), is one of the two most commonly used in DIF studies at ETS (Educational Testing Service) and was recommended by Howard Wainer (1989), a principal research scientist at that institution.

In the Mantel-Haenszel procedure, differences in item performance between the group members at each score level are weighted and then summed across score levels to obtain a common-odds ratio depicting the direction and degree of differential performance. The resulting common-odds ratio is logarithmically transformed to produce a delta value, which provides an estimate of the DIF where 0 represents the absence of DIF on an item, negative values indicate that the focal group members performed less well than the reference groups, and positive values indicate that the focal group performed better than the reference groups.

General Item Characteristics

Following the procedures of previous DIF studies (Bleistein & Wright, 1986; Dorans & Kulick, 1986; Freedle & Kostin, 1987; Schmitt & Bleistein, 1987), we examined the items for possible explanatory factors that could be correlated with the resulting DIF values. Five individuals, who were trained using techniques developed in previous studies (Freedle & Kostin, 1987; Lawrence, Curley, & McHale, 1988), conducted the item ratings.

Three characteristics functioned consistently across all of the 85 items in the verbal subtest:

- Subject matter content: (a) aesthetics or philosophy, (b) world of practical affairs, (c) science, (d) human relationships, (e) environment, (f) culture;
- position within the section;
- word abstractness.

Characteristics Unique To Each Item Type

We used item characteristics identified in previous studies as being related to DIF. In many instances the characteristics examined were identified through visual observation or statistical analyses rather than through theoretical considerations. The sentence completion items were rated on (a) whether outside knowledge about the content of the sentence would have been advantageous to the examinee and (b) the number of blanks in the item.

We classified the reading comprehension items according to the length of the corresponding passage and the type of information requested (e.g., main idea, inference, application), and we rated the analogy items for the presence of homographs, vertical relationships and semantic relationships. A multiple regression analysis was conducted to determine the extent to which each item characteristic accounted for the DIF values.

Results

The analysis of the 1988 *SAT* data identified noticeable DIF on only antonym item types and more specifically the earlier items in each section. Items with delta values greater than .60 or less than -.60 were in most instances statistically significant and viewed as exhibiting DIF. Nine items exhibited DIF in which the Hawai'i examinees were performing less well than the reference group, and on eight items Hawai'i examinees performed better than the reference group. Seven of the nine items showing negative DIF values which were less than or equal to -.60 were antonym items. The only item showing a DIF less than -1.00 was also an antonym item. There were eight items with DIF values greater than or equal to .60. Four of the items were from the reading comprehension section, and two items each were from the antonyms and analogy sections.

Distribution of DIF for the Four Verbal Item Types

Two items exhibited extreme DIF values (absolute value > 1.0). Antonyms constituted the only item type in which there were a large number of items on which the Hawai'i students performed differentially less well than the Mainland examinees did.

One item exhibited large DIF against Hawai'i examinees, an antonym item involving the word "spontaneous." The item found to have large DIF against the reference group was a reading comprehension item, which referred to a passage dealing with a view of Black literature and required the examinee to determine what the author suggested was an important quality of a character described in the passage.

Item Position—Common Characteristics

The only other characteristic which was comparable across all items in the verbal section dealt with item position within the section. Across all items in the test there was a tendency for the performance of Hawai'i examinees to improve (relatively) as they got further into each section.(see Table 1.) The relationship, however, was not statistically significant at the .05 level.

Table 1.
Pearson Product-Moment Correlations of Item Position and DIF

Section	r	Prob>t
Overall	.13	.23
Antonyms	.43	.03
Sentence Completion	-.17	.54
Analogies	.16	.49
Reading Comprehension	.04	.86

We found a statistically significant correlation among the antonym items (r=.43, p= 03), between the DIF values and item position indicating there was a tendency for the Hawai'i examinees to perform less well than the reference group on the early antonym items in each section and better than the reference group on the items later in each section.

Item Characteristics by Item Type

The results of regressing the DIF values on the item characteristics by item type showed no statistically significant effects at the .05 level. Individual items within specific item characteristic groups exhibited large DIF values but are not given as much attention as the mean DIF value for the entire group.

The items contained within each of the four types of items within the verbal section were classified according to a predetermined set of characteristics in the attempt to account for the differences in performance besides item type, item position, subject matter content, and degree of abstractness. Because several item characteristics were not applicable for all 85 items, detailed analyses were conducted within each of the four types of items.

Antonyms

The results of regressing the item characteristics of the antonym items on the DIF values are displayed in Table 2. Item position was the only statistically significant (α=.05) characteristic that had an effect on the DIF values (F=5.60, Prob>F=.03).

Table 2.
Simultaneous Regression of Antonym Item Characteristics on DIF Values.

Characteristic	Unique Sum of Squares	F	Prob>F
Overall Model	2.01	1.26	.32
Item Position	1.48	5.60	.03
Positive-Negative	.22	.41	.67
Personality Reference	.11	.42	.52
Type of Word	.20	.75	.40
Degree of Abstractness	.21	.79	.38

Overall $R^2 = .29$

Study 2. Experimental Treatment Addressing DIF

Study 2 was designed to address the performance discrepancies identified in the first part of the study and thus focused on antonym strategies and making students aware of their low performance on the easier items.

Sample

Thirty-seven Hawai'i students from two public high schools participated in Study 2. We selected the schools for their small but stable number of students being administered the test each year and the stability of their mean *SAT* scores over a 5-year period. The Hawai'i sample of students from the 1988 testing with *SAT-V* scores matched to the means of the two public high schools was used as the comparison group.

Procedure

A one-hour training session was developed to address the low performance of the Hawai'i students on the antonym items. The DIF patterns indicated that the low performance could possibly be due to an unfamiliarity with the item format or a lack of time being spent on the earlier items rather than the student's vocabulary knowledge.

We tentatively attributed the large, negative DIF scores on the early items in the antonym sections to an unfamiliarity effect in which the respondents use the first few items to get used to the format. Because DIF was present on only the early antonym items, which were located at the beginning of each of the two sections of the verbal subtest, we hypothesized that the students were anxious to move on to other parts in each section and thus did not spend enough time or put adequate amounts of concentration on the earlier items which contain less difficult words. Because the Hawai'i students had positive DIF scores on the later, more difficult items, they should have been able to respond to the earlier items at a higher accuracy rate. Inasmuch as the early items contained more readily recognizable words, the Hawai'i respondents may have rushed in selecting a response which was not the most nearly opposite or correct response.

The methods used in the training session were based on the premise of the unfamiliarity effect and a lack of adequate time spent on cognitively processing the words and evaluating the possible responses on the early items. The 1-hour session involved the following:

10-item pretest
Overview of the *SAT* and focusing on antonyms
Techniques in answering antonym items
Practice items
15-item posttest

The items for the pre- and post-tests were taken from the November 1988 version of the *SAT*. There items allowed us to compare the response patterns between the treatment students and the original sample (test familiarity was not a concern because the items had not been previously released).

In order to address the hypothesized unfamiliarity with the item format, we reviewed techniques for answering antonym items. The techniques were taken from the *Princeton Review* (Robinson & Katzman, 1991), *Peterson's Panic Plan for the SAT* (Carris, Crystal, & McQuade, 1990) and *10 SATs* (College Entrance Examination Board, 1988).

Analysis of Data

Because a minimum of approximately 300 students was needed to obtain reliable DIF values as had been obtained in Study 1, we did not use the Mantel-Haenszel procedure. Instead we used the *SAT* scoring formula, obtained by subtracting the number answered incorrectly divided by 4 from the number correct. The results from the pre- and post-tests, using the items from section two of the November 1988 *SAT* administered to the treatment students, were compared with the scores of students from the 1988 testing through independent groups *t*-tests.

Results

In order to test for a treatment effect we conducted two analyses: (1) a pre-post comparison among the treatment subjects, and (2) a comparison of the treatment results with a sample of the Hawai'i students' responses from the November 1988 administration.

The maximum raw score possible was 10 on the pretest and 15 on the posttest. Because the mean DIF of the antonym items from the November 1988 testing showed that Hawai'i students were doing less well than the reference group, the mean raw score for each group was obtained for validation. The results are presented in Table 3.

Table 3.
Mean Raw Score for Two Sections of Antonym Items

	Reference Group		Hawai'i Sample	
Section	Mean	SD	Mean	SD
10 Items	4.46	2.43	3.71	2.61
15 Items	7.08	3.54	6.03	3.75

The mean raw scores indicate that Hawai'i students are performing less well than the reference group by almost 1.8 raw score points on the 25 antonym items. According to the *SAT* raw-score-to-scaled-score conversion table the difference in raw score points corresponds to approximately 18 *SAT* formula score points. That amount is roughly the difference between Hawai'i and Mainland examinees on the *SAT-V* subtest as a whole. The treatment's goal was the improvement of the raw score performance of the Hawai'i students. The information presented in the treatment was based on the hypothesis that something other than vocabulary knowledge would improve the scores.

Obtaining a treatment group of students which would be representative of the Hawai'i examinees with a mean *SAT-V* of 408 would have been difficult. The lack of information about the scholastic verbal aptitude levels of students who have not yet been administered the test and access to student records made the attempt unfeasible. Instead we used the mean scores of the two schools willing to participate in the study to extract an equivalent range of students from the November 1988 testing.

We conducted statistical comparisons between the treatment groups and the Hawai'i sample of students from the November 1988 testing using independent groups *t*-tests. As shown in Table 4 the *t*-test comparisons between the Hawai'i and the treatment group pretest raw scores resulted in a nonsignificant *t* value (α=.05).

Table 4.
t -Test of Hawai'i Sample and Treatment Group Pretest

Group	Mean	N	*t*	Prob>t
10 Items (Pretest for Treatment Groups)				
Hawai'i Sample	2.26	85	.24	.81
Combined Treatment	2.34	37		

The *t*-test results between the treatment group posttest means and the 15-item antonym section of the November 1988 testing showed statistically significant differences. As shown in Table 5 the mean antonym raw score for the treatment group of 5.25 was statistically significantly ($p < .05$) larger than the Hawai'i sample mean of 3.85. The results indicate that while there were no statistically significant differences between the Hawai'i sample and the treatment pretest, the treatment group's posttest scores were statistically significantly higher than those of the comparison group.

Table 5.
t -Tests of Hawai'i Sample and Treatment Group Posttest Results

Group	Mean	N	*t*	Prob>*t*
15 Items (Posttest for Treatment Groups)				
Hawai'i Sample	3.85	85	2.44	.02
Combined Treatment	5.25	37		

Discussion

The DIF analysis established that Hawai'i students who take the *SAT* perform less well on the earlier antonym items than their Mainland U.S. counterparts of equal scholastic aptitude. The results also indicate that the roots of the low performance on the verbal subtest are possibly carelessness, rushing, and an unfamiliarity with the item format.

Pre- and posttest results for the one-hour treatment lend additional evidence for the factors which were hypothesized to be related to the low performance. More importantly the treatment shows how significant gains can be made in a relatively short period of time. The performance gap between Hawai'i and the national average would disappear if each student were to get slightly less than two more antonym items correct.

It is beyond the scope of this paper to deal with the current controversy about the validity of

standardized tests. One could certainly question any effort to raise the scores on such a test if the main value of the increase was simply political. On the other hand, this study has shown that the major reason for the relatively low verbal *SAT* scores of Hawai'i students essentially is not the relative lack of verbal aptitude.

Perhaps more important is the demonstration that systematic research methods can yield useful information that would otherwise be unavailable. Research procedures such as used in this study have the potential to identify areas where changes are likely to lead to improvements, whether they be in terms of better learning or better demonstration of what has been learned.

References

Bleistein, C. A., & Wright, D. (1986, April). *Assessment of unexpected differential item difficulty for Asian-American candidates on the Scholastic Aptitude Test.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Carris, J. D., Crystal, M. R., & McQuade, W. R. (1990). *Peterson's panic plan for the SAT.* Princeton, NJ: Peterson's Guides.

College Entrance Examination Board. (1988). *10 SATs.* New York: College Board Publications.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (Report 83-9). Princeton, NJ: Educational Testing Service.

Dorans, N., & Kulick, E. (1986). Demonstrating the utility of the Standardization approach to assessing unexpected differential item performance on the *Scholastic Aptitude Test. Journal of Educational Measurement, 23,* 355-368.

Freedle, R., & Kostin, I. (1987). *Semantic and structural factors affecting the performance of matched Black and White examinees on analogy items from the Scholastic Aptitude Test.* Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Hacketi, R. K., Holland, P., Pearlman, M., & Thayer, D. (1987). *Test construction manipulating score differences between Black and White examinees: Properties of the resulting tests* (Report 87-30). Princeton, NJ: Educational Testing Service.

Holland, P., & Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Lawrence, I. M., Curley, W. E., & McHale, F. (1988). *Differential item functioning of SAT-verbal reading subscore items for male and female examinees.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

Lord, F. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14,* 117-138.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Robinson, A., & Katzman, J. (1991). *The Princeton review: Cracking the system, the SAT and PSAT.* New York: Villard Books.

Schmitt, A. P. (1985). *Assessing unexpected differential item performance of Hispanic candidates on SAT form 3FSA08 and TSWE form E47* (Report 85-169). Princeton, NJ: Educational Testing Service.

Schmitt, A., & Bleistein, C. (1987). *Factors affecting differential item functioning for Black examinees on Scholastic Aptitude Test analogy items.* (Report 87-23). Princeton, NJ: Educational Testing Service.

Wainer, H. (1989). Personal communication.

Wright, D. (1983). *Assessing unexpected differential item performance of Oriental candidates and of White candidates for whom English is not the best language on SAT form 3FSA08 and TSWE form E47* (Report 85-123). Princeton, NJ: Educational Testing Service.